WPS5646

Policy Research Working Paper 5646

Good Countries or Good Projects?

Macro and Micro Correlates of World Bank Project Performance

Cevdet Denizer
Daniel Kaufmann
Aart Kraay

The World Bank
Development Research Group
Macroeconomics and Growth Team
May 2011

Policy Research Working Paper 5646

Abstract

The authors use data from more than 6,000 World Bank projects evaluated between 1983 and 2009 to investigate macro and micro correlates of project outcomes. They find that country-level "macro" measures of the quality of policies and institutions are very strongly correlated with project outcomes, confirming the importance of country-level performance for the effective use of aid resources. However, a striking feature of the data is that the success of individual development projects varies much more *within* countries than it does between countries. The authors assemble a large set of project-level "micro" correlates of project outcomes in an effort to explain some of this within-country variation. They find that measures of project size, the extent of project supervision, and evaluation lags are all significantly correlated with project outcomes, as are early-warning indicators that flag problematic projects during the implementation stage. They also find that measures of World Bank project task manager quality matter significantly for the ultimate outcome of projects. They discuss the implications of these findings for donor policies aimed at aid effectiveness.

This paper is a product of the Macroeconomics and Growth Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at akraay@worldbank.org.

Good Countries or Good Projects?
Macro and Micro Correlates of World Bank Project Performance

Cevdet Denizer (World Bank)
Daniel Kaufmann (Brookings Institution)
Aart Kraay (World Bank)

cdenizer@worldbank.org, dkaufmann@brookings.edu, akraay@worldbank.org. We are very grateful to Jaime Zaldivar and Kartheek Kandikuppa for their assistance in retrieving project-level data from the World Bank's databases, and to Martha Ainsworth, Antonella Bassani, Jaime Biderman, Jean-Jacques Dethier, Marguerite Duponchel, Homi Kharas, Alex McKenzie, Hari Prasad, Lant Pritchett, Veronika Penciakova, Luis Serven, Andrew Warner, and seminar participants at the World Bank and IMF for helpful discussions. Support from the Concessional Finance and Partnerships Vice Presidency of the World Bank is gratefully acknowledged. The views expressed here are the authors', and do not reflect those of the Brookings Institution, the World Bank, its Executive Directors, or the countries they represent.

1. Introduction

When is foreign aid effective in achieving its desired objectives? A vast empirical literature has sought to answer this question. Most of this literature has focused on the aggregate country-level impact of aid (typically on GDP growth), and has necessarily also focused on country-level factors that determine the growth effects of development assistance. [1] An important limitation of the aggregate country-level approach is that it has little to say about the very large variation in the success or failure of individual aid-financed development projects. In this paper, we contribute to the much smaller literature that seeks to provide empirical evidence on the factors contributing to the outcomes of individual projects, using data on over 6,000 World Bank-financed projects undertaken in 130 developing countries since the 1970s.

Our dependent variable consists of a subjective rating of the extent to which individual World Bank projects were able to attain their intended development objectives. These ratings are generated through internal World Bank project evaluation procedures, which we describe in more detail below. While we acknowledge upfront that these ratings are only imperfect indicators of actual project success or failure, we will for terminological convenience refer to these ratings as "project outcomes". We begin by documenting a set of very robust partial correlations between project outcomes and basic measures of country-level policy and institutional quality observed over the life of the project. This echoes other findings in the literature on macro-level determinants of aid effectiveness, which emphasize the role of country-level proxies for macroeconomic stability and the quality of policies and institutions in driving project outcomes. Most notably, we find a very strong partial correlation between the World Bank's Country Policy and Institutional Assessment (CPIA) ratings and project performance. While this result holds for all World Bank projects, it is particularly noteworthy in the context of projects financed by the International Development Association (IDA), which is the World Bank's fund for the poorest countries. IDA resources are allocated across countries using a Performance Based Allocation (PBA) process which directs more resources to countries that (a) are poorer, and (b) perform better on the CPIA ratings. It is therefore encouraging to see that this *ex ante* emphasis on country-level policy performance is reflected

[1] This line of research has produced a wide variety of conflicting results, to the point where Temple (2010) suggests that it "must be regarded as a work in progress". Recent assessments over the past decade range from cautiously optimistic Burnside and Dollar (2001), Clements, Radelet, Bhavnani (2004), Hansen and Tarp (2000), Minoiou and Reddy (2009) to ambivalent Roodman (2007) to skeptical and pessimistic Easterly and Levine (2004), Doucouliagos and Paldam (2008), and Rajan and Subramanian, (2008).

in *ex post* better project outcomes. Taken together, we find that such country-level variables can account for about 40 percent of the cross-country variation in project outcomes.

However, enthusiasm for this finding on the importance of country-level variables for project outcomes needs to be tempered by the observation that roughly 80 percent of the variation in project outcomes in our sample occurs across projects *within* countries, rather than *between* countries. While country-level variables explain a respectable fraction of the cross-country variation in average project outcomes, country-level variation comprises just 20 percent of the total variation in project outcomes. This basic observation suggests that there are very large returns to gathering and studying potential project-level correlates of project outcomes, which have largely been overlooked in the cross-country literature on aid effectiveness. The primary difficulty in doing so lies in obtaining such variables. To meet this challenge, we draw extensively on the World Bank's internal databases to extract three categories of such variables, all of which we discuss in greater detail below: (1) basic project characteristics such as the size and sector of the project, and the amount of resources devoted to the preparation and supervision of the project, (2) potential early-warning indicators of project success retrieved from the World Bank's institutional processes for monitoring and implementing active projects (the Implementation and Status Results (ISR) reports); and (3) information on the identity of the World Bank task manager associated with the project.

We find that several project-level variables, such as project size, project length, and whether the project was flagged as a "problem project" early in the life of the project, are important correlates of project-level outcomes. While these findings are encouraging, they too however are limited by the fact that these variables account for only a very small fraction of the within-country project-level variation in the data (on average around 6 percent of the within-country variation). Since the within-country variation in project performance is so large, however, this is still a substantial contribution to the overall explanatory power of our regressions: country-level variables account for about 8 percentage points of the total R-squared, while project-level variables account on average for about 5 percentage points of the total R-squared.

It is clear from these results that much more could be done to understand why project outcomes vary so much within countries. While we cannot provide a full accounting, in the final section of the paper we explore in a preliminary way one set of candidate explanations: the role of differences in task manager quality in explaining variation in project performance. We study this question in a reduced sample of projects where we have information on the identity of the task manager, and we also

have meaningful variation in project outcomes across both countries and task managers. Our headline finding here is that task manager fixed effects are at least as important as country fixed effects in accounting for the variation in project outcomes, suggesting a strong role for various task manager-specific characteristics in driving project outcomes.

The results in this paper have important implications for aid effectiveness in general, and for IDA in particular. The first is basic and not very new, though it is confirmed by the updated and expanded work in this paper: targeting aid to countries with better policies and institutions pays off, as rates of project success are significantly higher in countries with good policy, as measured by the CPIA ratings. However, the very large heterogeneity in project performance within countries suggests that policies to improve aid effectiveness could focus more on project-level factors as opposed to country-level factors. These include those that make individual projects difficult to restructure or cancel outright even after early indications of problems arise; those that contribute to project size and complexity; and those that underlie the large differences in project performance across task managers that we observe in the data.

The rest of this paper proceeds as follows. In the next section we review related literature that has also considered the World Bank project-level outcome data we work with here. In Section 3 we describe the project-level data in detail. Section 4 contains our main empirical results, and Section 5 discusses countries, projects, and task managers. Section 6 offers policy implications and conclusions.

2. Related Literature

This paper is not the first to study the correlates of individual World Bank project outcomes. In earlier contributions, Isham, Kaufmann and Pritchett (1997) and Isham and Kaufmann (1999) studied the determinants of project-level estimated ex-post economic rates of return. Both of these papers focused primarily on country-level factors affecting project returns, notably the role of democracy and civil liberties in the first, and the role of sound macroeconomic policies in the second. Subsequent papers have similarly focused on country-level determinants of project performance. For instance, Dollar and Levin (2005) estimate a series of cross-country regressions of country-average project success ratings on a set of country-level explanatory variables; they emphasize the role of differences in institutional quality in driving cross-country differences in average project performance. Guillamont and Laajaj (2006) focus on country-level volatility in accounting for project-level success, while Chauvet, Collier, and Duponchel (2010) emphasize country-level conflict measures. In addition Dreher, Klasen, Vreeland and Werker (2010) focus on the effect of political influence in the project approval decisions (as proxied

by a country-level variable capturing whether the country benefitting from the project was a member of the World Bank's Executive Board) on project outcomes. Finally, World Bank (2010, Chapter 7) studies the impact of trends in country-level macro variables, such as growth and market-oriented reforms, on variation over time in project-level Economic Rates of Return (ERRs).

While there is very large project-level variation in the data, only a handful of previous papers have sought to link this project-level variation in outcomes to project-level explanatory variables. Deininger, Squire, and Basu (1998) primarily focus on the effect that the volume of pre-existing country-level economic analysis has on the success of projects, but also contrast this with a project-level variable measuring the time spent by World Bank staff on project supervision. Dollar and Svensson (2000) focus on a small set of structural adjustment projects and investigate the role of both country-level political economy factors and a number of project-level factors, such as project preparation and supervision time and the number of conditions associated with the loan, in determining the ultimate success of structural adjustment operations. Kilby (2000) examines the role of staff supervision in determining project outcomes, but focuses on a set of interim outcome measures gathered over the course of project implementation, rather than the ex post outcome measures used in most other papers, including this one. Chauvet, Collier, and Fuster (2006) also emphasize supervision, and document the differential effect it has on project outcomes in countries with strong and weak governance. Pohl and Mihaljek (1998) focus less on project outcomes themselves, and more on the discrepancy between ex ante and ex post estimated economic rates of return at the project level.

Finally, while not focused on World Bank projects, our emphasis on the distinction between country-level and project-level correlates of project outcomes is shared with Khwaja (2009), who investigates the role of project-level and community-level characteristics in determining the success of individual small infrastructure projects undertaken in a set of communities in Northern Pakistan. He documents that community-level constraints to successful project performance can be alleviated by better design at the level of individual projects, thus enabling "good" projects in "bad" communities.

3. Institutional Background and Data

Institutional Background

In order to understand the data on project performance used in this paper, some institutional background is helpful. The lending activities of the World Bank are organized by project. For example, a project might consist of an agreement to build a particular piece of infrastructure, to fund teacher or

health worker training, to support a particular health intervention or a myriad of other potential development-oriented government actions that the World Bank finances. In some cases, projects simply take the form of budget support to receiving countries. A document describing the project is prepared by World Bank staff and includes a proposed amount of World Bank funding. An important ingredient in this initial document is a statement of the project's "development objective," which summarizes what the project intends to achieve. This development objective is important because the evaluation data we use rates the performance of the project relative to this objective. Once the project is approved by the Board of Executive Directors of the World Bank, it is implemented over several years, with the project spending financed by disbursements on loans provided by the World Bank.

On the World Bank side, each project is staffed by a project team and led by a task manager (formally known as a "task team leader"). At least twice a year, World Bank task managers are required to report on the status of the projects for which they are responsible, by completing an Implementation and Status Results (ISR) report (previously known as a Project Status Report (PSR)). As discussed below, these ISRs provide us with a rich set of project-level variables measured over the life of the project. Naturally, the World Bank's information systems also capture the annual disbursements on loans associated with each project over its duration. We use this information to construct disbursement rates on projects. Once a project is concluded, the responsible task manager produces an Implementation Completion Report (ICR) (previously known as a Project Completion Report (PCR)), which provides a comprehensive review of various dimensions of project outcomes. As we discuss further below, the reports form the basis of several of our outcome measures.[2]

All of the results that follow are representative only of the set of World Bank projects actually implemented over the past 30 years and for which we have detailed data. There should be no presumption that this set of World Bank projects is necessarily representative of all potential public investment projects in developing countries, and our results should be interpreted with this in mind. Indeed, the process by which projects are selected and implemented reflects a complex balancing of World Bank and recipient country interests and priorities, and therefore has implications for the broader relevance of our results. For instance, it is plausible that the World Bank has much more influence over the set of projects it finances in the poorest, aid dependent countries than those implemented in richer

[2] It is worth noting that projects are typically "complete" in the data once disbursements are completed and the implementation stage of the project is finished. Of course many projects are intended to continue delivering benefits for many years after this completion date. And by the same token, these benefits may not be apparent for many years after "completion" as well.

ones. The importance of country policies and institutions in accounting for the cross-country variation in project performance will be understated to the extent that the Bank uses its influence to select a small set of projects that are more likely to succeed in poor countries with weak institutional capacity, while financing "typical" projects in richer and/or better governed countries.

Project Outcome Data

Our primary project-level outcome variable is a subjective assessment of the extent to which a project met its stated "development objective". These project outcome assessments are available since 1972, when the Independent Evaluation Group (IEG) (previously known as the Operations Evaluation Department (OED)), was established within the World Bank. For projects evaluated prior to 1995, our main project outcome measure is a binary rating of satisfactory/unsatisfactory (IEGSAT). For projects evaluated since 1995, a six-point rating scale is used (*IEGRATE*). For the post-1995 period we use *IEGRATE*, and in addition we extend the binary *IEGSAT* over the entire sample period by recoding the top three categories of *IEGRATE* as "satisfactory".

These project outcome ratings come from three sources. As noted above, all completed projects since 1972 have a satisfactory/unsatisfactory rating drawn from the ICR. These can be thought of as an initial "self-evaluation" by World Bank staff and management of the project. These ratings are produced by the task manager, and are subject to review by World Bank management of the country/region where the project took place. After 1995, all such ICR-based evaluations were also subject to an additional layer of validation by IEG, based on available project documentation (these desk reviews were variously known as "Evaluation Summaries" or "Evaluation Memoranda", which we refer to as EVM). Finally, over the entire period since 1972, a sample of about 25 percent of projects completed each year are selected by IEG for a more detailed ex-post evaluation, known as "Project Performance Audit Reports" (PPAR). These typically occur several years after project completion – the mean lag in our core sample between project completion and these detailed IEG evaluations is 3.8 years, as opposed to 1.5 years for the first two types. In order to control for any effect these variations in

evaluation lags may have on outcomes, we will control for the time elapsed between project completion and project evaluation (*EVALLAG*) in the empirical work that follows.[3]

For our measure of project outcomes we take the outcome rating from the most detailed review available. That is, we first consider the IEG Project Performance Assessment Report rating for those projects where it is available. When it is not available, we take the ratings from the IEG desk reviews. For the remaining projects we use the outcomes as recorded in the ICR. In our core regression sample, we have 6,253 project outcomes, of which 2,022 ratings based on detailed IEG reviews, 2,934 ratings based on IEG desk reviews, and the remaining 1,297 ratings are based on Implementation Completion Reports alone.

There may naturally be questions regarding the credibility of these ratings. A basic concern might be that ICR-based ratings primarily reflect the view of the task manager, who may not be fully candid about the shortcomings of the projects. To test whether this concern is important, in Figure 1 we graph the average number of projects rated as "satisfactory" over time for each of the three types of evaluations (projects are organized here by year of evaluation). During the period up to 1995 we can compare the ICR-based reviews with the more detailed PPAR evaluations. This simple comparison shows little difference in the average rating across these two types of evaluations. During the period after 1995 there also do not appear to be very strong differences in average ratings between PPAR and EVM evaluations. More formally, in the empirical work that follows, we always include dummy variables for evaluation type to capture any mean differences in project outcome ratings across evaluation methods. Consistent with the evidence in Figure 1, these dummies rarely enter significantly.[4]

Another concern might be the credibility of the IEG evaluations themselves. On the one hand, several factors point to the plausibility of IEG ratings. The IEG is formally independent of the rest of the Bank's management and directly reports to the Board. Its review procedures are developed independently, its staff is experienced with evaluation issues, and has the ability to draw on cross-

[3] After 1993, several other subjective assessments of project outcomes are available, including a distinction between "overall borrower performance" and "overall bank performance". In the data, these other outcomes are extremely highly correlated with the overall project outcome ratings, and moreover are prepared by the same evaluator, and so it is unclear how much independent information they bring. For this reason we focus only on the overall outcome. In addition, for about half of projects in our sample estimated ex ante and ex post economic rate of return estimates are available. In subsequent drafts we will examine how our results carry over to this other, more quantitative, measure of the impact of the project.

[4] A separate issue is whether PPAR evaluations produced by IEG result in lower scores for the same project than the initial PCR evaluation completed by the task team leader. There is some evidence that this is the case unconditionally, when comparing projects for which *both* evaluations are available.

country and cross-project experience to inform project assessments and apply common standards. Moreover, since the 1990s most IEG evaluations have been public and the Group pays close attention to comments and criticisms of outside experts, civil society groups, and academia. On the other hand, IEG is primarily staffed by current and future Bank staff and there is some rotation in and out of IEG, although this turnover is considerably lower than in other parts of the Bank. There are also likely various informal channels of communication between IEG and World Bank staff which may affect the ratings process. While the full independence of IEG evaluations cannot be directly verified or contradicted, we can do no more than raise this as a potential question regarding the reliability of outcome measures.

A further qualification regarding these project outcomes ratings is that they are explicitly measured relative to the stated "development objective" of each project, rather than relative to some common standard across them. It could be the case that task managers set modest "development objectives" for projects undertaken in difficult countries, and set more ambitious ones in countries with good institutions, policies, and a good track record of implementing World Bank projects. This would understate the importance of country factors in determining the success of projects. At the same time, it could be the case that more experienced task managers set more modest development objectives, which are then more likely to be attained. This would in turn result in an overstatement of the importance of task manager effects in driving project success.[5]

Our overall impression is that while the evaluation outcome data described here are far from perfect, arguably they meaningfully capture the experience and insights over the years of many World Bank staff on how well projects have fared. Of course, even well-measured individual project outcomes will not be perfectly informative about the overall aggregate development impact of aid, as there are there may be complementarities between projects, as well as the potential scope for aid-financed spending crowding out other sorts of public spending. Nevertheless, while surely there is considerable remaining measurement error in the outcome measures, it is still useful to investigate a range of

[5] However, the Bank's review process has some safeguards to prevent this. First, during the concept review stage of project preparation the teams are routinely questioned about the "realism" of project objectives. In addition, for much of the sample period, World Bank staff from the Quality Assurance Group (QAG) reviewed a sample of projects in "real time" and advised teams and management about various aspects of proposed projects, including their development objectives. While there is no formal common standard for development objectives, there are formal mechanisms in place that attempt to ensure that project objectives are both feasible and ambitious enough to attain the broader development objectives of a the recipient country.

country-level and project-level factors that are associated with these outcomes. We discuss these factors next.

Country-Level Correlates of Project Performance

We consider a small set of core country-level variables that have been identified in the literature as important correlates of project outcomes. We first include the logarithm of one plus the inflation rate as a crude proxy for macroeconomic stability (*LNINFAV*). We also measure the quality of country-level policies and institutions using the Country Policy and Institutional Assessment (CPIA), ratings of the World Bank (*CPIAAV*).[6] These variables have been emphasized in several earlier papers (notably Isham and Kaufmann (1999) and Dollar and Levin (2005)), and are intended to capture the effect of country-level policies and institutions on project performance. We also include real per capita GDP growth as a crude proxy for macroeconomic shocks (*DYAV*). Finally, we consider the role of civil liberties and political rights as country-level correlates of project performance, as emphasized by Isham, Kaufmann and Pritchett (1997), for the particular economic rate of return outcome measure. We measure these using the sum of the Freedom House scores of civil liberties and political rights (*FRHAV)*.

Our unit of observation is a project, for which the execution and implementation typically lasts several years – the median length of a project in our sample is 6 years and 10 percent of projects last 9 years or more. For each project, we calculate the annual average of each of these country-level correlates of performance over the life of the project (from approval to completion). We will use these project-life averages as explanatory variables for project performance in the empirical work below.

Project-Level Correlates of Project Performance

Our first set of project-level variables captures basic project characteristics. We use the logarithm (in millions USD) of the total amount of World Bank lending committed to each project as a basic measure of size (*LNSIZE*). Larger projects tend to be complex, with multiple components, tranches and counterparts who implement individual components, all of which adds complexity to project implementation. We also have information on the start and end dates of each project, which

[6] A potential concern regarding the CPIA measure is that it also is scored by Bank staff, leading to a possible mechanical relationship between country-level CPIA scores and project-level outcome ratings. This however is less of a concern given that we average all the country-level variables over the life of the project, while the outcome is measured at or after the end of the project. In any case, as a robustness check we also used the Worldwide Governance Indicators Rule of Law measure (www.govindicators.org) and found very similar results (not reported for reasons of space).

together gives us its length. We use the logarithm of this (in years) as a measure of project duration (*LNLENGTH*). We also have information on the distribution of projects across economic sectors (summarized in Table 1). We use a set of sector dummies to capture any sectoral variation in average project outcomes. Finally, we have data on preparation and supervision costs for each project. We express this as a fraction of the total size of the project, and use the logarithm of this in our regressions (*LNPREPCOST* and *LNSUPERCOST*).

A second set of project-level variables is obtained from the ISR process of monitoring and eventual evaluation of projects. We have retrieved data from the end-of-fiscal-year ISR for every project, and for each year during the life of the project. This provides us with information reported by the task manager on a variety of interim measures of project performance, on an annual basis over the life of the project. Each year, the implementation status of the project is rated relative to the ultimate development objective, and if this rating is unsatisfactory, the project is flagged as a "problem project". In addition, task managers indicate with a series of 12 flags whether there are concerns about specific dimensions of project performance, including problems with financial management, compliance with safeguard, quality of monitoring and evaluation, legal issues, etc. If three or more of these flags are raised at any one point in time, the project is identified as a "potential problem project". Even if a project is flagged as a potential problem project, country management units have the authority to override this classification using a "golden flag", on which we also have data.

While in principle the flag data are a rich source of information on leading indicators of project performance, they need to be treated with some caution. While some of the flags are automatically triggered (for example, by objectively-measured disbursement delays, or by lags between project approval and the start of work on the project), the decision to raise others is at the discretion of task managers, who for natural reasons may be reluctant to do so. This could be due to optimism about the ultimate outcome of the project, or reputational concerns on the part of the task manager and/or counterparts. Indeed, a perennial concern for Bank management has been the frequency of projects exhibiting "disconnect" – projects that were rated as satisfactory throughout the implementation process but were then ultimately rated as unsatisfactory upon completion. Despite these caveats, these flags are an extremely important set of candidate predictors of project success since they are routinely generated and are readily available to World Bank decision makers who can act on them to improve the ultimate outcome of the project.

For each project we construct a set of dummy variables indicating whether each flag was raised in the first half of the project implementation period, measured in calendar years. For example, for a project lasting 6 years from approval to completion, we construct dummy variables indicating whether each of the flags was raised in the first three years of the project. We then investigate whether these "early warning" flags are related to the eventual outcome of the project. Creating this lag between the measured flags and the completion of the project is important for two reasons. First, our primary interest in these flags is as potential leading indicators of eventual project outcomes. In particular, we would like to investigate whether flagging a project as a "problem" or "potential problem" early in its life creates opportunities or incentives to take remedial steps to turn the project around – this is after all the point of having a process for monitoring projects over the course of implementation. Second, we would like to avoid any mechanical link between the ISR flags and the ultimate project outcome rating as captured in the ICR or subsequent IEG review. Consider for example the "problem project" flag, which is supposed to be raised in the ISR process if the project is not making satisfactory progress towards its development objective. This criterion is very similar to the ultimate project outcome rating in the ICR, which, as discussed above, captures the extent to which the project was able to meet its development objective.

Another potential predictor of project performance comes from annual disbursement flows on individual projects. For each project approved since 1985, we have data on the actual amount disbursed annually on each project. We use this data to construct a measure of disbursement delay as a potential leading indicator of project success. To do so, we need to compare actual disbursements with expected or planned disbursements. Each Bank project includes a disbursement schedule which provides an initial estimate of expected disbursements over its life. The performance of actual disbursements relative to these plans are monitored in the ISRs, and considerable institutional attention is paid to deviations from them, particularly to disbursement lags as this would suggest slower-than-expected project implementation. A priori it is unclear what the correlation of slower-than-average disbursements will be with project outcomes. On the one hand, slow disbursements may signal careful project implementation with strict fiduciary safeguards, which may lead to better project outcomes. On the other hand, disbursement delays may signal projects that were prepared and approved with excessive haste, and were unprepared for timely implementation. In this case, slow disbursements might be associated with worse outcomes.

The problem with these data on disbursement lags however is that initial disbursement projections are tentative and often revised over the life of the project. Thus, they are a kind of a "moving target" if we want to measure what actual disbursements delays are. In light of this, we instead measure disbursement delays relative to typical disbursement rates for projects approved in the same region, year, and sector, drawing from Kraay (2010). In particular, for each project we construct a typical disbursement profile based on the actual average annual disbursement rate for all projects in the same sector, approval year, and geographical region. For each year of the project, we then calculate the difference between the cumulative actual annual disbursement rate and the cumulative predicted annual disbursement rate. Finally, we take this cumulative disbursement lag over the first half of the project, and consider it as a potential leading indicator of project success (*DISBLAGH1*).

Table 2 contains summary statistics on the main variables of interest. These are reported for two samples that are defined based on various data availability constraints. Our first sample begins in 1983 (i.e. with projects evaluated in 1983), and contains 6,253 subsequent project evaluations (after eliminating 548 observations with missing data on the relevant right-hand-side variables). In this sample, the outcome rating is a binary satisfactory/unsatisfactory, and we have data available on the indicated basic project characteristics as well as the ISR flags data. There are a further 810 project evaluations between 1972 and 1982 but we do not include them as the ISR flag data, as well as the preparation and supervision costs data, is not available from this earlier period. Our second sample consists of 3,887 evaluations performed since 1995. For this sample, we use the available six-point scale for the outcomes, and also a richer set of control variables including the disbursement lags and a limited set task manager variables, that we describe in more detail below.

4. Results

Country-Level Correlates of Project Performance

We begin in Table 3 by documenting the role of country-level variables in explaining the heterogeneity in project outcomes. The left panel of Table 3 reports results for all World Bank projects, and the right panel for IDA projects separately. Within each panel, the first two columns report results for the full post-1983 sample, and the remaining two for the post-1995 sub-sample.[7] Throughout the

[7] Throughout the paper we will report results from simple OLS regressions, even though the outcome variable is binary in the first sample, and discrete on a six-point scale in the second. This is purely for pragmatic reasons. While an ordered multinomial choice model is in principle more appropriate for the second sample, when the number of categories is large, the value-added of recognizing explicitly the discrete nature of the dependent

paper, the unit of observation is the project, and the standard errors are clustered at the country-by-evaluation-year level to allow for correlations in errors across projects evaluated in the same year-by-country cell.

In all specifications, we include a set of basic controls for the characteristics of the type of evaluation on which the dependent variable is based. As discussed above, in the post-1983 sample we combine data from three types of evaluations, and therefore include dummy variables for each type in our regressions (i.e. we include a dummy for detailed Project Performance Assessment Review (*PARDUM*) and lighter IEG desk reviews (*EVMDUM*), with task-manager-generated Implementation Completion Report assessments as the omitted category. In the post-1995 sample, we have no ICR-based outcomes, and so we include only the PPAR dummy (with EVM as the omitted category). The dummies for evaluation type are not statistically significant, indicating that there is no evidence of average differences in scores for the two more detailed evaluation types relative to the benchmark task manager-reported evaluations. Interestingly, there is a negative relationship between reported project outcomes and the delay between project completion and evaluation (*EVALLAG*). This effect is highly significant in the full sample of projects, and remains negative but insignificant among IDA projects. One interpretation is that letting more time elapse between project completion and evaluation allows problems with the project to become more apparent. It is striking as well that this result holds even when we include a dummy for PPAR evaluations (*PARDUM*) – indicating that it is not the case that the *EVALLAG* variable is simply capturing "tougher" assessments of the detailed PPAR reviews, which typically occur well after the project is completed. In fact, conditional on the evaluation lag, there is no evidence that PPAR evaluations result in significantly lower scores.

Turning to the country-level variables, we find that growth and policy performance, as measured by the CPIA, enter significantly and with the expected signs. In particular, the CPIA variable (*CPIAAV*) is very significant and the coefficient is meaningfully large: in columns (2) and (4) for example, a one-point improvement in the CPIA (on a six-point scale) implies on average about a 0.5 point improvement in the six-point project outcome rating. In terms of standard deviations, a one-standard-deviation improvement in country-level policy performance is associated with a 0.23 standard deviation improvement in the six-point project outcome rating. This strong partial correlation between country

---

variable is small, and comes at the expense of greater difficulty in interpreting the coefficients. We therefore use OLS in this sample. And to make results more comparable with the larger sample we also use OLS rather than probit. However, re-estimating our core specifications using a probit makes virtually no difference for the patterns of size and significance of the estimated slope coefficients.

policy performance and project outcomes can be interpreted as validating the importance of CPIA rating in determining the allocation of IDA resources across countries.[8]

The partial correlation between country-level aggregate growth (*DYAV*) and project performance is also highly significant, although the standardized magnitude of the effect is somewhat smaller than that of policy performance. A one-standard-deviation increase in aggregate GDP growth over the life of the project (i.e. an increase in growth of 3.2%) is associated with a 0.12-standard-deviation improvement in project performance (i.e. an increase of 0.15 points on a six-point scale). The results for inflation (*LNINFLAV*), our proxy for macroeconomic stability, are more mixed. In the post-1983 sample, inflation enters negatively and highly significantly, indicating that macroeconomic instability is strongly associated with worse project performance. This effect largely disappears in the post-1995 sample. The results are at least partially driven by the fact that episodes of very high inflation, reflecting extreme macroeconomic mismanagement, are much less common in the post-1995 sample. In all specifications the Freedom House measure (*FRHAV*) has a positive link with project performance, as previously documented by Isham, Kaufmann and Pritchett (1997) using ex post economic rates of return as a measure of project outcomes. However, this effect is significant at conventional levels only in the first column.

Finally, it is worth noting that although these country-level variables are generally significant correlates of project performance, they jointly have rather modest explanatory power; the R-squareds of the regressions in Table 3 range from 0.07 to 0.09. However, as we will see in the next section, these country-level variables account for a much greater share of the cross-country variation in project performance. But, since this country-level variation accounts for only around 20 percent of the total variation in project outcomes, the overall explanatory power of country-level variables remains moderate. This motivates our exploration of the explanatory power of a range of project-level variables for project outcomes, to which we turn next.

Project-Level Correlates of Project Performance

In Table 4 we consider the first set of basic project characteristics that might be associated with project outcomes. To conserve space, we do not report the estimated coefficients of the evaluation characteristics and country-level variables considered in Table 3, though all of these variables are

[8] However, it is somewhat ambiguous who should get the "credit" for the positive association between CPIA scores and project performance. See for example Wane (2004) who notes that countries with capable governments might be better positioned to resist pressures from donors to take on poorly-designed projects.

included in the specifications shown in Table 4. There is evidence that larger projects, in dollar terms, are less likely to result in satisfactory outcomes -- the variable *LNSIZE* enters negatively and significantly in three of the four specifications, suggesting that project complexity (as proxied by size) is associated with worse project performance. There is also a very strong negative partial correlation between project length (*PROJLENGTH*) and project outcomes, with longer projects associated with significantly worse outcomes in all four specifications. However, the interpretation of this partial correlation is complicated by the fact that we have information only on actual project length, not initially-planned project length. As a result, it is difficult to assign any causal interpretation to this correlation. It could simply be the case that some projects, especially the larger ones, "go wrong" and then take longer to complete than originally intended.

A similar concern clouds the interpretation of the negative partial correlation between project supervision costs (*LNSUPERCOST*) and project outcomes. A plausible interpretation of this result is that when a project starts to go wrong, more resources are devoted to supervision in an effort to turn around the project. However, these efforts do not always succeed, and so the data show a negative partial correlation between supervision and project outcomes. Unfortunately, we do not have data on the distribution of supervision costs over the life of the project – given this information one could investigate whether more intensive early supervision is able to turn around problematic projects.

A much more surprising finding is the consistently negative partial correlation between project preparation costs (*LNPREPCOST*) and project outcomes, though the results are not statistically significant at conventional levels. This indicates that on average eventual outcomes are less likely to be satisfactory in projects where more money is spent on preparation. This finding is particularly surprising because preparation costs are by definition incurred before the project begins. One interpretation of this finding is that initial costs signal "high-risk" projects, possibly with less recipient country ownership, that are (a) more likely to require intensive preparation and (b) more likely to ultimately receive unsatisfactory ratings. Finally, we include a set of five dummies indicating major project sectors (Agriculture and Rural Development, Energy and Mining, Transport, Education, and Economic Policy, with the omitted category consisting of an aggregate of the remaining sectors noted in Table 1). A strikingly consistent pattern is that projects in the Transport and Education sectors consistently perform better on average than projects in other sectors.

In Table 5, we introduce "real-time" project monitoring data from the ISRs, specifically the annual series of "problem" and "potential problem" flags over the life of each project. As discussed

above, we construct variables indicating whether these flags were raised in the first half of project execution (measured in calendar years). In addition, we restrict attention to projects that lasted at least four years from approval to completion, in order to ensure that there is a meaningful lag between the raising of any flags and ultimate project outcome rating. We add these flag indicators to the specifications from Table 4. As before, to conserve space we do not report the coefficients of all the variables considered in the previous two tables, although they are included in the regressions.

We first introduce dummy variables indicating whether a project was flagged as a "problem" or "potential problem" during the first half of its life (*PROBLEMH1* and *POTPROBLEMH1*, in columns (1), (3), (5), and (7)). In all four cases, these variables enter negatively and very significantly, indicating a strong partial correlation between the early flagging of a project and its ultimate outcome. This finding suggests that even when early warning flags are raised through the ISRs, it is difficult to turn around problematic projects in order to achieve satisfactory outcomes.

Another way of seeing this is to consider directly the persistence of problem project flags. Consider the post-1983 sample in column (1). Of the 4,560 projects in this sample, 1,146 or about 25 percent of the total were flagged as problem projects in the first half of their lives. Two-thirds of these (772/1146) were also flagged as problem projects in the second half of their lives. On the other hand, one-third of projects initially flagged as problematic were turned around during their second half and became non-problematic. This persistence is also found in the final project outcomes that comprise our main dependent variable of interest. Unconditionally, projects that are flagged as a problem in the first half of their implementation period have only a 56 percent chance of yielding satisfactory results, while projects that are not flagged in their first half have a 75 percent chance of turning out satisfactorily. Interestingly, projects that are deemed problematic during their first half, but not during their second (indicating that initial problems have been resolved), have an 83 percent change of ultimately being deemed satisfactory.

In the post-1995 sample, we also have data on disbursement delays during the first half of each project. Interestingly, there is no evidence that disbursement delays are significantly correlated with project outcomes. As discussed above, this may reflect the balance of two opposing forces. On the one hand, faster-than-average disbursements could indicate a well-prepared project that is being implemented successfully. On the other hand, it could also signal overly-fast disbursements relative to the project's absorptive capacity, eventually leading to an unsatisfactory project rating. Qualitatively however, in three of the four specifications the disbursement lags the variable enters positively if not

significantly, suggesting that the latter effect dominates. While not statistically significant, the sign of this correlation calls into question the institutional emphasis placed by the World Bank on disbursement delays.

As noted above, the potential problem flag is raised only if three or more subsidiary flags are raised during the ISR process. In columns (2), (4), (6) and (8) we investigate in more detail which of these individual flags matter most for project outcomes Slightly more than half of the 52 reported coefficients (13 flags times four specifications) turn out to be negative, consistent with the idea that early adverse flags are associated with worse project outcomes. However, only in a few cases are these negative partial correlations statistically significant. One striking example is the flag indicating concerns with the monitoring and evaluation components of projects (*FLAGMONEVALH1*). This finding highlights the crucial importance of strong monitoring mechanisms in the design and implementation of projects. Among IDA projects, the Country Record flag (*FLAGCTRYRCRDH1*) is also significantly negative. This is a country-level flag that is triggered if disbursements on unsatisfactory projects represent more than 20 percent of total disbursements to the country over the previous five years. This suggests strong inertia in the quality of projects over time, even conditional on the contemporaneous overall policy environment as captured by the CPIA.

A somewhat puzzling finding from the disaggregated flags is that the flag indicating concerns about the overall country environment (*FLAGCTRYENVH1*) is significantly positively correlated with project outcomes. This could be due to the fact that we have already controlled for the overall policy environment using CPIA. A final noteworthy observation about the disaggregated flags is that the "golden flag" (*FLAGGOLDENH1*) is not significantly positively correlated with project outcomes. Recall that it can be used at the discretion of Bank management to override other flags. These results could reflect the balance of two opposing forces. On the one hand, if a golden flag is only deployed if there are additional concerns with project implementation, then this selection effect would lead to lower average outcome ratings. On the other hand, a golden flag may be used to signal that identified implementation problems are not sufficiently severe to lead to unsatisfactory project outcomes.

5. Countries, Projects, and Task Managers

Good Countries or Good Projects?

A key contribution of this paper is its consideration of a broad set of project-level correlates of project success. This contrasts with much of the existing literature, which has primarily emphasized

country-level correlates of project outcomes. We now take stock of the relative contributions of these two types of variables in accounting for the variation in project performance. As noted in the introduction, a key feature of the data is that *within* country variation in project-level outcomes is greater than *between* country variation. We document this stylized fact using the set of all projects for which we have an evaluation outcome as well as a full set of control variables (i.e. the set of 4,560 projects included in the regression sample in Column 1 of Table 5). For each year between 1985 and 2005 we select the set of projects that were active in that year, and regress the eventual outcome of these on a set of country fixed effects.[9] The R-squared from this regression captures the share of variance in project outcomes that is due to country-level factors, and is shown in the first column of Table 6. This *between*-country variance, which we refer to as the "macro" variation in project outcomes, accounts for a surprisingly small share of total variance, averaging only 20 percent. The remaining 80 percent of variation in project outcomes occurs across projects *within* the same country, i.e. the "micro" variation.

Next we document the relative importance of country and project-level variables in accounting for the "macro" and "micro" variation in project performance. In each year, we regress project outcomes on the full set of country- and project-level variables included in Column (1) of Table 5, and again retrieve the R-squared. The results are reported in the third column of Table 6 for each year. This R-squared captures the share of the variation in project outcomes that is jointly accounted for by the full set of country-level and project-level variables included in the regression. The year-by-year R-squareds are fairly modest, averaging around 12 percent, and of course are similar to those in Table 5 which pool all the annual cross-sections.

This R-squared can be decomposed into a country-level and a project-level component as follows:

$$R^2 = R^2_{MACRO} V_{MACRO} + R^2_{MICRO}(1 - V_{MACRO}) \quad (1)$$

The between-country or "macro" R-squared, R^2_{MACRO}, is the R-squared that would be obtained from a pure cross-country regression of country-average project outcomes on country-level variables. This

[9] The entries in this table stop in 2005 since in order to enter in our regression sample, projects active in 2005 also need to be evaluated by 2009. Going beyond 2005 results in too small a sample of projects to permit meaningful inference.

captures how well country-level variables account for the cross-country variation in project performance, and is reported in the fourth column of Table 6. This between-country R-squared is quite respectable, averaging 40 percent. This indicates that the parsimonious set of country-level variables used here accounts for a close to half of the average cross-country differences in project performance. However, as we have seen, cross-country variation in project performance accounts for only about 20 percent of the variation in overall project outcomes. So, overall country-level "macro" variables can explain just 8 percent (40 percent times 20 percent) of the total variation in project performance.

As the first two columns of Table 6 indicate, differences between projects within countries account for the bulk of variation in project outcomes, and thus to account for this variation it is necessary to look at project-level explanatory variables. The within-country "micro" R-squared, R^2_{MICRO}, captures the explanatory power of the project-level variables in accounting for the micro" variation in project outcomes. While many of the project-level variables we have considered have been individually significant, collectively they account for only a very modest share of the "micro" variation in the data, as the "micro" R-squareds average just 6 percent. However, explaining just 6 percent of the 80 percent of the variation in project outcomes within countries nevertheless contributes importantly to the fit of the regression. Overall, the project level variables contribute around 5 percent (6 percent times 80 percent) of overall variation in project performance.

Good Countries or Good Task Managers?

As noted above, a great deal of the variation in project outcomes occurs across projects within individual countries. So far, we have investigated the contribution of a range of project characteristics and early-warning signals. As shown in the previous section, our efforts in this respect have been only modestly successful, in that the project-level variables explain only a small part of the very substantial project-level variation in outcomes. In this section we investigate another potentially-important correlate of project performance, the identity of the task manager responsible for the project. We have obtained data on the staff identification number of the World Bank task manager at the time of completion, for a set of 3,921 projects evaluated since 1995. An important limitation of this data, however, is that task managers can change over the life of a project, but we only have information about the identity of the task manager at the end of each project.

In order to have a meaningful comparison of the relative importance of task manager characteristics and country characteristics in accounting for project outcomes, we need to restrict

attention to projects that (a) were managed by task managers who also managed several other projects, and in more than one country, and (b) occur in countries where there are multiple projects. This ensures that our sample will have meaningful variation across both countries and task managers. We first focus only on projects whose task manager is associated with at least four other evaluated projects (for a total of at least five projects per task manager), and whose task manager also has projects in more than one country. This reduces the sample greatly to 930 projects, scattered across 118 countries, many of which have just a few projects in the sample. To allow for meaningful within-country variation in project performance we further restrict attention to only those projects occurring in countries where at least four other projects can be evaluated, for a total of at least five projects per country. This limits the sample to 790 projects in 68 countries. This final sample contains 150 distinct task managers. Given our particular interest in distinguishing between task manager and country effects, it makes sense to work with this greatly restricted sample, even though it is probably too small to deliver reliable inferences on the relative role of other project and country characteristics in driving project outcomes that we examined earlier in the paper using much larger samples.

In the first two columns of Table 7, we simply regress project outcome ratings on a full set of task manager dummies, and then on a full set of country dummies. The results of this simple exercise are striking: task manager effects account for 32 percent of the variation in project outcomes, while pure country effects account for only 19 percent of the variation.[10] These results suggests that in our sample, differences in the characteristics of task managers matter much more for project outcomes than do differences in the characteristics of the countries where they are implemented.

While this stylized fact is striking, it is also difficult to interpret because it ignores any correlation between task manager characteristics and the countries in which they work. If for example higher-quality task managers tend to work in countries with good policy and good CPIA scores, then distinguishing between country effects and task manager effects is difficult. We investigate this further in columns (3) and (4), where we introduce variables intended to capture some of the task manager and country characteristics that matter for project outcome. Specifically, we introduce a very crude proxy for "task manager quality" for each project, consisting of the average rating of all of the other projects

[10] One might worry that the simple comparison of R-squareds is misleading because there are more task manager dummies (150) than country dummies (68). However, the relative importance of task manager effects persists if we instead compare adjusted R-squareds, which are 17 percent for the task manager dummies versus 11 percent for the country dummies. One might also worry that task managers tend to specialize in particular sectors, and that task manager effects are simply picking up some of the sectoral differences in project outcomes noted previously. However, we obtain very similar results when sector fixed effects are included in the regression.

managed by the same task manager, excluding the project in question (*TMQUAL*). To capture country effects, we consider, as before, the average over the life of the project of the country-level CPIA score.

We first regress project outcomes on task manager quality and a full set of country dummies (column (3)). Task manager quality is again highly significant, indicating that even within countries, projects associated with higher quality task managers are more likely to be rated as satisfactory. In Column (4) we ask the symmetric question, regressing project outcomes on the CPIA score and a full set of task manager dummies. The CPIA is also highly significant, indicating that the projects of a given task manager are more likely to be rated as satisfactory in countries with good CPIA scores.

Finally, we include both CPIA scores and task manager quality, and find that both are very significant predictors of project success. The relative magnitude of these effects is interesting. The standard deviation of *TMQUAL* in the sample is 0.77, and so a one-standard-deviation improvement in task manager quality leads to an improvement in the project outcome rating of 0.26 (on a six-point scale, or about 0.2 standard deviations of the outcome measure). For *CPIAAV* the standard deviation is 0.59, and so a one-standard-deviation improvement in country-level policy performance implies an improvement of 0.25 in the project outcome measure. This suggests that the magnitude of the standardized impact of task manager quality and policy quality in the recipient country on project outcomes is roughly the same.

6. Interpretation, Implications, and Conclusions

In this paper, we analyze the correlates of project outcomes for a very large sample of World Bank projects since the early 1980s. In our analysis, we have distinguished between country-level correlates of country-average project performance, and project-level correlates of project-level variation in project outcomes *within* countries. This distinction is important as the within-country share of the variation in project outcomes is large, at around 80 percent. Consistent with existing literature, we find that country-level variables, most notably the CPIA measure of policy and institutional quality, are robust partial correlates of project performance. This basic finding underscores the importance of continued country-level selectivity in aid allocation. In the case of projects funded by IDA, these results can be seen as a validation of IDA's Performance Based Allocation system which emphasizes "macro" country-level measures of policy and institutional quality in determining the cross-country allocation of aid.

However, a great deal of the very substantial variation in project outcomes within countries remains unexplained by country-level factors, and our analysis points to the importance of project-level or "micro" factors in explaining some of this variation. Several of our specific findings have important policy implications in this respect, and we highlight a few here. A quite robust empirical finding is that the longer the time elapsed between project completion and evaluation, the less likely the project will be rated as satisfactory. One interpretation of this is that early evaluations are overoptimistic and that the true impact of a project is apparent only over time. This in turn suggests that the quality of evaluations could be improved simply by allowing more time to elapse between project completion and project evaluation, so that the actual outcomes of the project have more time to become apparent. This also points to the importance of creating incentives that reward the long-term impact of projects rather than simply their satisfactory completion. Indeed, the impacts for projects in sectors such as education and health usually show over time and it may therefore be more appropriate to evaluate these types of projects much later than those implemented in other sectors.

A second important result is that there are some strong early-warning indicators of project outcome ratings. One is simply project preparation costs; there is a robust partial correlation between higher preparation costs and eventual low project outcome ratings. These high preparation costs might reflect undue initial project complexity, or limited country ownership, or various other factors, that cannot be overcome despite considerable resources being devoted to preparation. Another is the set of early warning flags, several of which are strongly significant predictors of ultimate project outcomes. This finding holds for the broader “problem project” flag, and notably also for the more specific “monitoring and evaluation” flag. This suggests that the overall rate of satisfactory World Bank projects could be improved if incentives to significantly restructure or simply cancel problem projects at the implementation stage were strengthened, and by increased emphasis on monitoring and evaluation over the life of the project.

Another finding is that project size matters. Since large projects usually contain more components they tend to be more complex. This would mean that project size and design are important correlates of project performance, and that avoiding factors associated with undue project complexity could improve development results. A fourth finding with important policy implications is that task manager characteristics are important and have quantitatively large and significant impacts on project performance. Indeed, our simplest results indicate that task manager fixed effects are at least as important as country fixed effects in accounting for variation in project outcomes. This points to the

importance of internal policies to develop and propagate task manager skills in order to ensure better project outcomes.

The final policy implication comes from the humbling fact that even after accounting for a wide range of micro and macro variables, much of the variation in project performance remains unexplained. After all, in our core specifications we can account for only about 12 percent of the variation in measured project outcomes. Some of this is surely just measurement error, pointing to the importance of developing more robust tools for capturing project performance. But at the same time, much of this variation is likely to be real and it reflects a wide range of as-yet-unmeasured factors at both the country and project levels. Developing empirical proxies for these other factors, and thinking creatively about how to use them to design selectivity at both the country and project levels based on such factors, will ultimately help to improve overall aid effectiveness, for all aid donors, and for IDA in particular.

In particular, our findings suggest that cross-country aid allocation mechanisms, such as IDA's Performance Based Allocation system, could be complemented by project level mechanisms as well, consistent with the results-based orientation of IDA. This is consistent with Gelb (2010) who notes that finding a reasonable way of allocating at least of a portion of aid based on project-level indicators, especially in fragile states, could provide greater incentives for countries to ensure project success. In this way, successful projects in low institutional development environments could be scaled up and aid effectiveness could be improved. As a matter of fact, as more IDA countries "graduate", it is likely that IDA will be left with more countries with poor CPIA scores, and developing a way of allocating aid utilizing project level indicators is likely to become a necessary complement to country-level aid allocation policies.

References

Burnside, Craig, and David Dollar (2000). "Aid, Policies, and Growth". American Economic Review. 90(4):847-868.

Chauvet, Lisa, Paul Collier, and Margeurite Duponchel (2010). "What Explains Aid Project Success in Post-Conflict Situations?". World Bank Policy Research Working Paper No. 5418.

Chauvet, Lisa, Paul Collier, and Andreas Fuster (2006). "Supervision and Project Performance: A Principal-Agent Approach". Manuscript, DIAL.

Clemens, Michael, Steven Radelet, and Rikhil Bhavnani (2004). "Counting Chickens When They Hatch: The Short-Term Effect of Aid on Growth. Center for Global Development Working Paper No. 44.

Deininger, Klaus, Lyn Squire, and Swati Basu (1998). "Does Economic Analysis Improve the Quality of Foreign Assistance?". World Bank Economic Review. 12(3):385-418.

Dollar, David and Jakob Svensson (2000). "What Explains the Success and Failure of Structural Adjustment Programs?". The Economic Journal. 110():894-917.

Dollar, David and Victoria Levin (2005). "Sowing and Reaping: Institutional Quality and Project Outcomes in Developing Countries". World Bank Policy Research Working Paper No. 3524.

Doucouliagos, H., and Paldam, M. (2009). "The Aid Effectiveness Literature: The Sad Results of 40 Years of Research". Journal of Economic Surveys. 23(3): 433-461.

Dreher, Axel, Stephan Klasen, James Raymond Vreeland, and Eric Werker (2010). "The Costs of Favouritism: Is Politically-Driven Aid Less Effective?. CESIFO Working Paper No. 2993.

Easterly, William, Ross Levine, and David Roodman (2004). "Aid, policies, and growth: A Comment". American Economic Review. 94(3):774-780.

Gelb, Alan (2010). "How Can Donors Create Incentives for Results and Flexibility for Fragile States? A Proposal for IDA". Center for Global Development, Working Paper No. 227.

Guillaumont, Patrick and Rachid Laajaj (2006). "When Instability Increases the Effectiveness of Aid Projects". World Bank Policy Research Working Paper No. 4034.

Hansen, Henrik, and Finn Tarp (2000). "Aid Effectiveness Disputed". Journal of International Development. 12: 375-398

Isham, Jonathan and Daniel Kaufmann (1999). "The Forgotten Rationale for Policy Reform: The Productivity of Investment Projects". Quarterly Journal of Economics. 114(1):149-184

Isham, Jonathan, Daniel Kaufmann and Lant Pritchett (1997). "Civil Liberties, Democracy, and the Performance of Government Projects". World Bank Economic Review. 11(2): 219-242.

Khwaja, Asim Ijaz (2009). "Can Good Projects Succeed in Bad Communities?" Journal of Public Economics. 93: 899-916.

Kilby, Christopher (2000). "Supervision and Performance: The Case of World Bank Projects". Journal of Development Economics. 62: 233-259.

Kraay, Aart (2010). "How Large is the Government Spending Multiplier? Evidence from World Bank Lending. World Bank Policy Research Department Working Paper No. 5500.

Minoiu, C., and Reddy, Sanjay (2009). "Development Aid and Economic Growth: A Positive Long Term Relation". IMF Working Paper 09/118.

Pohl, Gerhard and Dubravko Mihaljek (1998). "Project Evaluation and Uncertainty in Practice: A Statistical Analysis of Rate-of-Return Divergences in 1015 World Bank Projects". World Bank Economic Review. 6(2): 255-257.

Rajan, Raghuram, and Arvind Subramanian (2008). "Aid and growth; What does the cross-country evidence really show?" Review of Economics and Statistics. 90(4):643-665.

Roodman, David (2007). "The anarchy of numbers: Aid, Development, and Cross-Country Empirics". World Bank Economic Review. 21(2): 255-277

Temple, Jonathan (2010). "Aid and Conditionality" in Handbook of Development Economics, Rodrik. D., and Rosenzweig, M. (eds)., Volume 5., Elsevier BV. (4415-4523).

Wane, Waly (2004). "The Quality of Foreign Aid: Country Selectivity or Donor Incentives?". World Bank Policy Research Department No. 3325.

World Bank (2010). "Cost-Benefit Analysis in World Bank Projects". Independent Evaluation Group.

Figure 1: Average Satisfactory Ratings Over Time, By Type of Evaluation

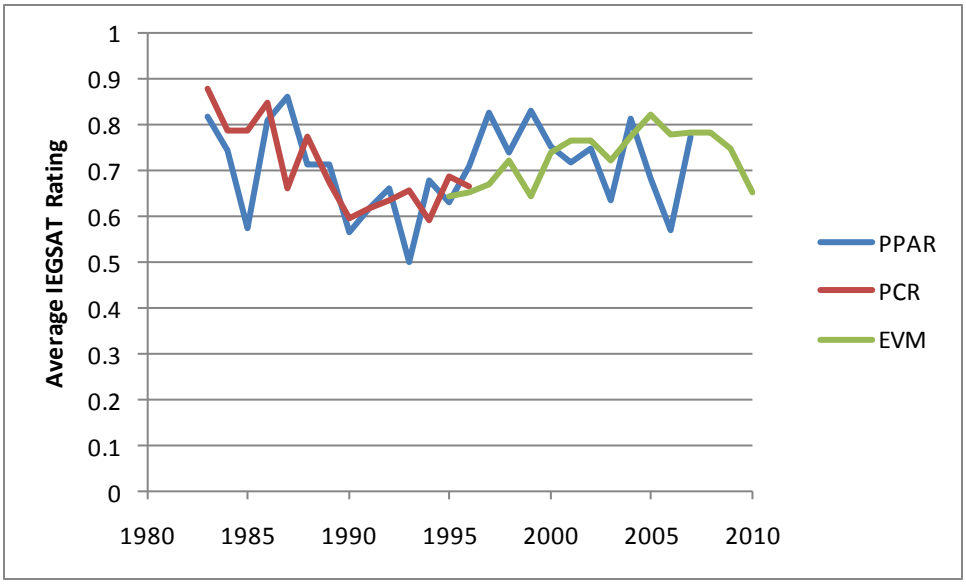

Table 1: Distribution of Projects Across Sectors

Sector	Number of Projects	Percent of Total
Agriculture and Rural Development	1,439	23.01
Energy and Mining	859	13.74
Transport	703	11.24
Education	560	8.96
Economic Policy	384	6.14
Other Sectors:	2308	36.93
Public Sector Governance	350	5.6
Financial and Private Sector Developmen	320	5.12
Health, Nutrition and Population	319	5.1
Urban Development	318	5.09
Water	314	5.02
Financial Sector	219	3.5
Social Protection	196	3.13
Environment	106	1.7
Global Information/Communications Techn	93	1.49
Social Development	30	0.48
Poverty Reduction	26	0.42
Private Sector Development	16	0.26
Gender and Development	1	0.02

Table 2: Summary Statistics

		1983-2009			1995-2009	
		(6253 Observations)			(3887 Observations)	
Description	**Code**	**Mean**	**Std. Dev.**		**Mean**	**Std. Dev.**
Outcome Variables						
IEG Satisfactory (1) / Unsatisfactory (0) Rating	iegsat	0.712	0.453			
IEG 6-point rating (1=bad, 6=good)	iegrate				4.066	1.251
Evaluation Characteristics						
Years from completion to evaluation	evallag	2.296	1.927		1.840	1.758
Dummy=1 for detailed IEG PPAR Evaluation	pardum	0.323	0.468		0.236	0.425
Dummy=1 for IEG ES/EVM Review	evmdum	0.469	0.499		0.736	0.441
Macro Variables (averaged over life of project)						
logarithm of (1+inflation)	lninfav	0.184	0.321		0.153	0.280
real GDP growth	dyav	0.020	0.032		0.025	0.032
CPIA score (1=bad, 6=good)	cpiaav	3.615	0.676		3.644	0.562
Freedom House (Average of CivLib and PolRight)	frhav	2.861	1.492		2.998	1.494
Basic Project Characteristics						
Logarithm of Original Committment	lnsize	3.605	1.215		3.781	1.229
Years from approval to completion	projlength	5.907	2.426		5.922	2.591
Logarithm of preparation costs/commitment	lnprepcost	-5.268	1.260		-5.102	1.204
Logarithm of supervision costs/commitment	lnsupercost	-4.941	1.369		-4.810	1.428
Early Warning Indicators (Measured in First Half of Project, Subsample of Projects >= 6 Years Long)						
Problem project flag	problemH1	0.251	0.434		0.279	0.449
Potential problem project flag	potproblemH1	0.155	0.362		0.214	0.410
Project effectiveness delay flag	flageffdelayH1	0.196	0.397		0.274	0.446
Counterpart funding problem flag	flagcntrprfundH	0.010	0.098		0.014	0.117
Financial management problem flag	flagfinmgmtH1	0.006	0.077		0.008	0.090
Safeguard problem flag	flagsafegrdH1	0.019	0.138		0.028	0.164
Monitoring and evaluation problem flag	flagmonevalH1	0.123	0.328		0.166	0.372
Legal covenant problem flag	flaglegcovH1	0.083	0.276		0.118	0.322
Country environment problem flag	flagctryenvH1	0.264	0.441		0.368	0.482
Procurement problem flag	flagprocmntH1	0.125	0.331		0.175	0.380
Project management problem flag	flagprojmgmtH1	0.187	0.390		0.203	0.402
Country record problem flag	flagctryrcrdH1	0.280	0.449		0.386	0.487
Long term project risk flag	flagltriskH1	0.130	0.336		0.180	0.384
Slow disbursement flag	flagslowdisbH1	0.149	0.356		0.210	0.407
Golden flag	flaggoldenH1	0.013	0.112		0.018	0.134
Disbursement lag	disblagH1				-0.010	0.178

Table 3: Country-Level Variables and Project Outcomes

	(1)	(2)	(3)	(4)
Sample	All Projects	All Projects	IDA Only	IDA Only
Time Period	1983-2009	1995-2009	1983-2009	1995-2009
Dependent Variable	iegsat	iegrate	iegsat	iegrate
evallag	-0.0170***	-0.0603***	-0.00967	-0.0361
	(-3.78)	(-3.24)	(-1.52)	(-1.56)
pardum	-0.00864	0.116*	-0.0343	0.0204
	(-0.45)	(1.66)	(-1.15)	(0.21)
evmdum	-0.0179		-0.0318	
	(-0.98)		(-1.15)	
lninfav	-0.0691**	0.125	-0.0763**	0.0226
	(-2.18)	(1.31)	(-2.17)	(0.24)
dyav	1.757***	4.718***	1.886***	4.952***
	(8.01)	(6.25)	(6.66)	(5.47)
cpiaav	0.0927***	0.516***	0.103***	0.532***
	(8.50)	(11.27)	(6.29)	(7.29)
frhav	0.00965**	0.0204	0.00372	-0.00798
	(2.31)	(1.35)	(0.55)	(-0.34)
_cons	0.377***	2.070***	0.351***	2.067***
	(8.61)	(12.68)	(5.69)	(8.24)
N	6253	3887	2816	1904
R-sq	0.071	0.093	0.068	0.085
t statistics in parentheses				
* p<0.10, ** p<0.05, *** p<0.01				

Table 4: Project-Level Variables and Project Outcomes

	(1)	(2)	(3)	(4)
Sample	All Projects	All Projects	IDA Only	IDA Only
Evaluation Characteristics?	Yes	Yes	Yes	Yes
Country Variables?	Yes	Yes	Yes	Yes
Time Period	1983-2009	1995-2009	1983-2009	1995-2009
Dependent Variable	iegsat	iegrate	iegsat	iegrate
lnsize	-0.0448***	-0.0760**	-0.0353**	-0.0708
	(-4.50)	(-2.01)	(-2.36)	(-1.37)
projlength	-0.0111***	-0.0418***	-0.0200***	-0.0560***
	(-3.12)	(-3.69)	(-4.03)	(-3.63)
lnprepcost	-0.00859	-0.0216	-0.0164	-0.0616
	(-1.13)	(-0.71)	(-1.42)	(-1.56)
lnsupercost	-0.0458***	-0.0797**	-0.0501***	-0.0907*
	(-4.59)	(-2.28)	(-3.50)	(-1.94)
agrurdev	-0.0155	0.0424	0.00765	0.0611
	(-0.87)	(0.73)	(0.33)	(0.79)
energymining	-0.00285	-0.122	0.00830	-0.0569
	(-0.14)	(-1.55)	(0.25)	(-0.50)
transport	0.0978***	0.435***	0.137***	0.523***
	(5.49)	(7.17)	(4.76)	(5.49)
education	0.0963***	0.226***	0.114***	0.191**
	(4.77)	(3.53)	(4.01)	(2.18)
econpolicy	-0.0461*	-0.0198	-0.1000***	-0.224**
	(-1.70)	(-0.23)	(-2.59)	(-1.96)
_cons	0.351***	2.201***	0.258***	1.975***
	(5.85)	(11.46)	(3.00)	(6.82)
N	6253	3887	2816	1904
R-sq	0.094	0.120	0.105	0.123
t statistics in parentheses				
* p<0.10, ** p<0.05, *** p<0.01				

Table 5: Early Warning Indicators

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	All Projects	All Projects	All Projects	All Projects	IDA Only	IDA Only	IDA Only	IDA Only
Evaluation Characteristics?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country Variables?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Basic Project Variables?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sample Period	1983-2009	1983-2009	1995-2009	1995-2009	1983-2009	1983-2009	1995-2009	1995-2009
Dependent Variable	iegsat	iegsat	iegrate	iegrate	iegsat	iegsat	iegrate	iegrate
problemH1	-0.143***	-0.0996***	-0.387***	-0.153**	-0.128***	-0.0954***	-0.298***	-0.0845
	(-8.13)	(-4.82)	(-7.02)	(-2.23)	(-5.14)	(-3.11)	(-4.00)	(-0.88)
potproblemH1	-0.0418**		-0.171***		-0.0839***		-0.250***	
	(-1.98)		(-2.83)		(-2.90)		(-2.90)	
flageffdelayH1		-0.0148		-0.0540		-0.0223		-0.0361
		(-0.82)		(-1.04)		(-0.81)		(-0.45)
flagcntrprtfundH1		-0.151*		-0.246		-0.112		-0.211
		(-1.86)		(-1.32)		(-1.03)		(-0.94)
flagfinmgmtH1		0.0241		0.0277		0.121		0.205
		(0.28)		(0.14)		(1.19)		(0.79)
flagsafegrdH1		0.0259		0.170		0.0710		0.400**
		(0.55)		(1.23)		(1.05)		(1.99)
flagmonevalH1		-0.189***		-0.675***		-0.140***		-0.458***
		(-7.17)		(-8.78)		(-3.70)		(-4.22)
flaglegcovH1		0.0530*		0.0736		0.0618		0.130
		(1.78)		(0.92)		(1.45)		(1.12)
flagctryenvH1		0.0444**		0.115**		0.0319		0.0836
		(2.46)		(2.21)		(1.33)		(1.16)
flagprocmntH1		0.00319		-0.0825		0.0106		-0.117
		(0.13)		(-1.28)		(0.31)		(-1.28)
flagprojmgmtH1		-0.0372*		-0.0937		-0.0434		-0.229**
		(-1.69)		(-1.36)		(-1.49)		(-2.42)
flagctryrcrdH1		0.0185		-0.0170		-0.0422*		-0.172**
		(1.01)		(-0.33)		(-1.65)		(-2.49)
flagltriskH1		-0.0104		-0.0843		-0.000884		-0.0373
		(-0.51)		(-1.45)		(-0.03)		(-0.45)
flagslowdisbH1		0.0221		0.0594		0.0193		0.114
		(1.14)		(1.09)		(0.67)		(1.36)
flaggoldenH1		0.0110		0.193		0.0136		0.260
		(0.20)		(1.28)		(0.17)		(1.26)
disblagH1			0.182	0.155			0.00432	-0.0128
			(1.40)	(1.23)			(0.02)	(-0.07)
_cons	0.360***	0.298***	2.399***	2.155***	0.235**	0.193*	2.366***	2.201***
	(4.74)	(3.82)	(10.70)	(9.07)	(2.25)	(1.80)	(6.70)	(5.74)
N	4560	4560	3022	3022	2130	2130	1484	1484
R-sq	0.106	0.122	0.139	0.174	0.117	0.126	0.140	0.166
t statistics in parentheses								
* p<0.10, ** p<0.05, *** p<0.01								

Table 6: Variance in Project Outcomes Between and Within Countries

	Variation in Dependent Variable			R-Squared			Contribution to R-Squared			Number of
Year	Macro	Micro		Total	Macro	Micro		Macro	Micro		Projects
1985	0.162	0.838		0.144	0.577	0.060		0.09352	0.05048		1249
1986	0.159	0.841		0.133	0.564	0.052		0.08964	0.04336		1266
1987	0.19	0.81		0.125	0.498	0.038		0.094625	0.030375		1273
1988	0.186	0.814		0.124	0.463	0.047		0.086136	0.037864		1256
1989	0.209	0.791		0.141	0.457	0.057		0.095612	0.045388		1296
1990	0.219	0.781		0.14	0.458	0.051		0.10032	0.03968		1338
1991	0.217	0.783		0.135	0.478	0.040		0.103685	0.031315		1366
1992	0.226	0.774		0.128	0.448	0.034		0.101304	0.026696		1386
1993	0.21	0.79		0.127	0.443	0.043		0.093018	0.033982		1373
1994	0.199	0.801		0.119	0.389	0.052		0.077422	0.041578		1386
1995	0.182	0.818		0.105	0.371	0.046		0.06744	0.03756		1383
1996	0.172	0.828		0.107	0.309	0.065		0.053231	0.053769		1416
1997	0.179	0.821		0.11	0.309	0.067		0.05529	0.05471		1405
1998	0.178	0.822		0.119	0.371	0.064		0.066113	0.052887		1400
1999	0.17	0.83		0.12	0.379	0.067		0.0644	0.0556		1346
2000	0.159	0.841		0.125	0.395	0.074		0.06275	0.06225		1291
2001	0.17	0.83		0.122	0.375	0.070		0.063762	0.058238		1178
2002	0.191	0.809		0.109	0.296	0.065		0.056459	0.052541		1051
2003	0.219	0.781		0.11	0.285	0.061		0.06236	0.04764		856
2004	0.279	0.721		0.113	0.266	0.054		0.074103	0.038897		641
2005	0.334	0.666		0.115	0.218	0.063		0.072925	0.042075		438
Average	0.200	0.800		0.122	0.398	0.056		0.078	0.045		

Table 7: Task Manager and Country Effects, 1995-2009

	(1)	(2)	(3)	(4)	(5)
Project Sample	All Projects	All Projects	All Projects	All Projects	All Projects
Sample Period					
Task Manager Dummies	Y	N	N	Y	N
Country Dummies	N	Y	Y	N	N
Dependent Variable	iegrate	iegrate	iegrate	iegrate	iegrate
tmqual			0.210***		0.340***
			(3.40)		(6.30)
cpiaav				0.464***	0.418***
				(4.67)	(5.89)
_cons	4.500***	3.600***	2.682***	2.496***	1.280***
	(9.98)	(7.07)	(4.68)	(4.04)	(4.15)
N	790	790	790	790	790
R-sq	0.323	0.188	0.201	0.346	0.109
t statistics in parentheses					
="* p<0.10	** p<0.05	*** p<0.01"			